# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-263514
(43)Date of publication of application : 11.10.1996

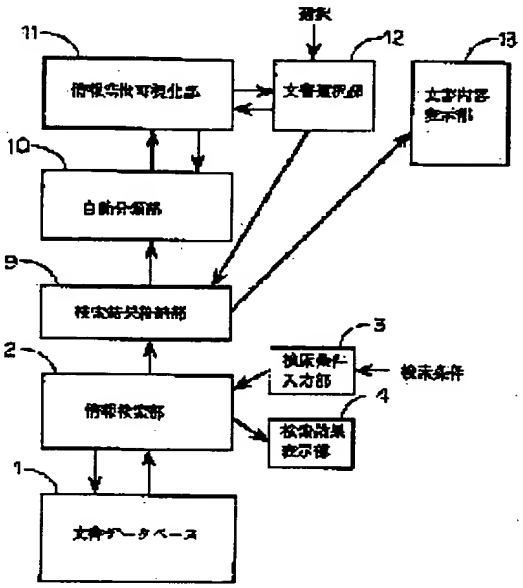
(51)Int.Cl. G06F 17/30
G06F 7/24

(21)Application number : 07-068160
(22)Date of filing : 27.03.1995

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : ARITA HIDEKAZU
YASUI TERUMASA
TSUDAKA SHINICHIRO

## (54) METHOD FOR AUTOMATIC CLASSIFICATION OF DOCUMENT, METHOD FOR VISUALIZATION OF INFORMATION SPACE, AND INFORMATION RETRIEVAL SYSTEM

(57)Abstract:
PURPOSE: To obtain the information retrieval system which can easily obtain a retrieval key word and functions as a bottom-up type sending support system.
CONSTITUTION: This system is provided with a retrieval result storage part 9 in which a set of documents retrieved from a document data base 1 by an information retrieval part 9 under retrieval conditions from a retrieval condition input part 3 is stored, an automatic classification part 10 which automatically classifies the documents in the document set, an information space visualization part 11 which visualizes the information space of the automatically classified document set, a document selection part 12 which calls an automatically classified document by specifying the visualized two-dimensional position and selects one of plural documents when there are plural documents, and a document content display part 13 which takes the contents of the selected document out of the retrieval storage part 9 and displays them.




LEGAL STATUS

[Date of request for examination] 01.07.1999
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number] 3385297
[Date of registration] 27.12.2002
[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－263514

(43)公開日　平成８年(1996)10月11日

| (51)Int.Cl.[6] | 識別記号 | 庁内整理番号 | ＦＩ | | 技術表示箇所 |
|---|---|---|---|---|---|
| Ｇ０６Ｆ　17/30 | | 9194－5L | Ｇ０６Ｆ　15/403 | ３５０Ｚ | |
| 7/24 | | | 7/24 | Ｌ | |
| | | 9194－5L | 15/403 | ３４０Ｂ | |
| | | 9194－5L | | ３５０Ｃ | |

審査請求　未請求　請求項の数６　ＯＬ　（全 18 頁）

(21)出願番号　特願平7－68160

(22)出願日　平成７年(1995)３月27日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号

(72)発明者　有田　英一
尼崎市塚口本町八丁目１番１号　三菱電機株式会社中央研究所内

(72)発明者　安井　照昌
尼崎市塚口本町八丁目１番１号　三菱電機株式会社中央研究所内

(72)発明者　津高　新一郎
尼崎市塚口本町八丁目１番１号　三菱電機株式会社中央研究所内

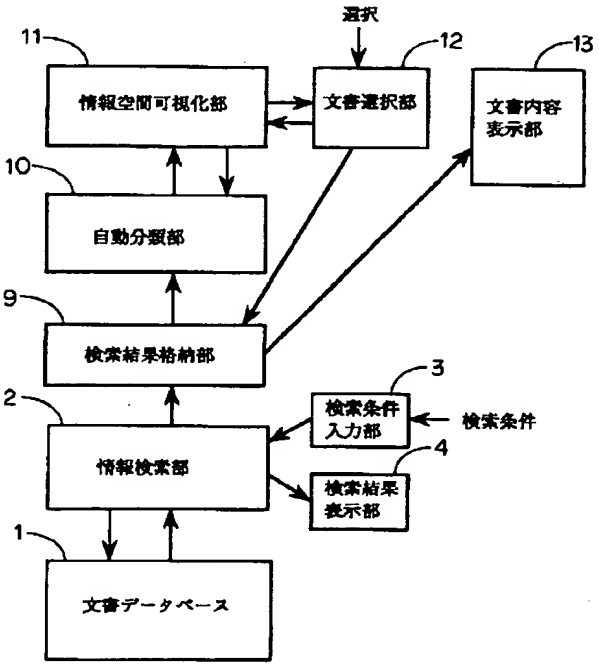
(74)代理人　弁理士　田澤　博昭　（外２名）

(54)【発明の名称】　文書の自動分類方法、および情報空間の可視化方法、ならびに情報検索システム

(57)【要約】

【目的】　検索キーワードが容易に得られ、ボトムアップ型の発送支援システムとして機能する情報検索システムを得る。

【構成】　検索条件入力部３からの検索条件に従って情報検索部２が文書データベース１より検索した文書集合を格納する検索結果格納部９、その文書集合を対象に文書を自動分類する自動分類部１０、自動分類された文書集合の情報空間を視覚化する情報空間可視化部１１、視覚化された２次元の位置を指定することにより自動分類された文書の呼び出しを行い、それが複数ある場合にはその中より１つを選択する文書選択部１２、および選択された文書の内容を検索結果格納部より取り出して表示する文書内容表示部１３を設けたもの。



1

【特許請求の範囲】

【請求項１】　文書集合が与えられた時、その文書集合中の各文書を、その内容に応じて自動的に分類する文書の自動分類方法において、前記文書集合に含まれる、単語や句、節などの意味のある文字列による語句の中から、一定の条件で選んだ語句を構成要素として、文書集合中のある文書を、それが含む語句に対応する値をその語句の出現頻度をもとに定めた語句ベクトルとして表現し、前記文書の語句ベクトルと２次元に配置されたセルに対応する語句ベクトルの距離を計算して、その距離が最小のものをその文書が所属する仮のセルとし、当該セルの語句ベクトルの要素の値をその文書の語句ベクトルの要素の値に近付けるとともに、そのセルの近傍のセルの語句ベクトルの要素の値も、その文書の語句ベクトルに対する近傍の度合いに応じて減じて近付けることを、前記文書集合に含まれる文書について一定回数、もしくは収束するまで実行し、その後、各セルの語句ベクトルと文書の語句ベクトルとの距離を計算して、その距離が最小のセルをその文書が所属する本来のセルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断することを特徴とする文書の自動分類方法。

【請求項２】　前記文書が所属する仮のセルが決まった後に、近傍のセルの語句ベクトルを文書の語句ベクトルに近付ける際、前記文書集合での語句の出現の割合に依存して、前記近傍のセルの語句ベクトルを文書の語句ベクトルに近付ける割合を変化させることにより、文書のクラスタの階層性と文書クラスタの所属するセルの領域の包含関係を対応させることを特徴とする請求項１に記載の文書の自動分類方法。

【請求項３】　請求項１または２に記載の文書の自動分類方法によって分類された文書集合の情報空間を可視化する情報空間の可視化方法において、セルに対応する語句ベクトルのうちの、値の大きなものから一定数、もしくは所定の閾値以上の値を持つものを、そのセルを代表する語句として、その文字列の表示を行うことを特徴とする情報空間の可視化方法。

【請求項４】　隣り合ったセルの境界線の属性を、セルに対応する語句ベクトルの距離に応じて変化させて表示することを特徴とする請求項３に記載の情報空間の可視化方法。

【請求項５】　複数の文書が格納された文書データベースと、前記文書データベースより検索条件に該当する文書を検索抽出する情報検索部と、前記情報検索部に検索条件を入力する検索条件入力部と、前記情報検索部による検索結果を出力表示する検索結果表示部とを有する情報検索システムにおいて、前記情報検索部にて検索された文書集合を格納する検索結果格納部と、前記検索結果格納部に格納された文書集合を対象に、請求項１または２に記載した文書の自動分類方法によって文書の自動分類を行う自動分類部と、前記自動分類部にて分類された

2

文書集合の情報空間を、請求項３または４に記載した情報空間の可視化方法によって視覚化する情報空間可視化部と、前記情報空間可視化部にて視覚化された２次元の位置を指定することによって前記自動分類された文書の選択を行い、それが複数ある場合にはその中の所定数の文書を選択する文書選択部と、前記文書選択部にて選択された文書の内容を前記検索結果格納部より取り出して表示する文書内容表示部を備えたことを特徴とする情報検索システム。

【請求項６】　前記検索結果格納部に格納された文書集合の各文書の特定部分を切り出して格納する文書分類選択格納部を設け、前記自動分類部が、前記文書の特定部分を入力テキストとして文書の自動分類を行うものであり、前記情報空間可視化部が、前記文書の特定部分を入力テキストとして自動分類された情報空間の可視化を行うものであることを特徴とする請求項５に記載の情報検索システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】この発明は、文書情報を自動的に分類する文書の自動分類方法、および分類された情報空間を可視化する情報空間の可視化方法、ならびに可視化された情報空間を参照して情報の検索を行う情報検索システムに関し、特に、内容の似た文書が近くに配置されるように２次元のセルに配置し、そのセルを代表する語句を表示して、文書情報の情報空間を一覧表示するとともに、文書データを簡単に検索できるようにして、ボトムアップ型の発想支援システムとして機能する情報検索システムに関するものである。

【０００２】

【従来の技術】図１３は従来の情報検索システムの機能構成を示す構成図である。図において、１は多量の文書情報が格納されている文書データベースであり、２はこの文書データベース１より検索条件に該当した文書データを検索抽出する情報検索部である。また、３はこの情報検索部２にキーワードの論理和や論理積などによる検索条件を入力する検索条件入力部であり、４は情報検索部２より出力される、検索条件に該当した文書の数や、必要に応じて該当する文書の見出しなどの検索結果を出力表示する検索結果表示部である。

【０００３】次に動作について説明する。現在実用化されている特許や科学技術文献などの文書情報を検索する場合、まず、あらかじめ規定されているシソーラスに記載された統制キーワードや、主として文書内に含まれている語句である自由キーワードの論理積や論理和等による検索式を作成し、それを検索条件として検索条件入力部３より情報検索部２に入力する。情報検索部２は文書データベース１を検索して、入力された検索条件に該当する文書を抽出し、当該検索条件に合致した文書の数や、必要に応じて検索された文書のタイトルや概要など

の検索結果を検索結果表示部４に出力表示する。

【０００４】なお、このようなこの発明に関連する従来のベクトルの自動分類方法について記載された文献としては、例えば「プロシーディングス　オブ　ザ　アイ・トリプル・イー（Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　Ｔｈｅ　ＩＥＥＥ）」の第７８巻第９号（１９９０年９月）の１４６４～１４８０ページに掲載された論文「ザ、セルフ・オーガナイジング　マップ（Ｔｈｅ　Ｓｅｌｆ－Ｏｒｇａｎｉｚｉｎｇ　Ｍａｐ）」などがある。

【０００５】

【発明が解決しようとする課題】従来の情報検索システムは以上のように構成されているので、検索条件に該当する文書が文書データベース全体の中でどのような位置を占めるのか明らかでないため、検索された文書が適切なものであるか否かの判断が困難であり、情報検索結果が文書単位でリストとしてしか得られず、内容については順番に本文を参照していく必要があって、内容の似たものを一括して見るのが困難であるばかりか、文書データベースの全体概要がわからず、文書があらかじめ定められた分類形態を基準に分類されていたとしても、分類のクラスタの相互関係が不明であり、さらに文書データベースにどのような自由キーワードがあるのかわからないなどの問題点があった。

【０００６】この発明は上記のような問題点を解消するためになされたもので、文書を自動分類して内容の近さを反映した２次元のセルとして配置し、各セルには分類を代表するキーワードを表示することによって文書データベースの全体構成を把握できるようにし、それを利用して検索キーワードが容易に得られる情報検索システムや、ボトムアップ型の発想支援システムとして機能する情報検索システム、さらには、それに用いられる文書の自動分類方法および情報空間の可視化方法を得ることを目的とする。

【０００７】

【課題を解決するための手段】請求項１に記載の発明に係る文書の自動分類方法は、文書の語句ベクトルとセルの語句ベクトルの距離を計算して、それが最小となるものをその文書の仮の所属セルとし、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近付けるとともに、そのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて減じて近付ける処理を、所定回数もしくは収束するまで繰り返し、その後、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断するものである。

【０００８】また、請求項２に記載の発明に係る文書の自動分類方法は、近傍のセルの語句ベクトルを文書の語句ベクトルに近付けるとき、語句の文書集合での出現割合に依存して近付ける割合を変化させ、文書のクラスタの階層性と文書クラスタの所属するセルの領域の包含関係を対応させるものである。



【０００９】また、請求項３に記載の発明に係る情報空間の可視化方法は、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句を、そのセルを代表する語句とし、当該語句の文字列を表示して、上記文書の自動分類方法によって分類された文書集合の情報空間を可視化するものである。

【００１０】また、請求項４に記載の発明に係る情報空間の可視化方法は、隣接するセルに対応する語句ベクトルの距離に応じて、そのセル間の境界線の属性を変えて表示するものである。

【００１１】また、請求項５に記載の発明に係る情報検索システムは、情報検索部にて検索された文書集合を格納する検索結果格納部、その文書集合を対象に、請求項１または２に記載した文書の自動分類方法によって文書の自動分類を行う自動分類部、分類された文書集合の情報空間を、請求項３または４に記載した情報空間の可視化方法によって視覚化する情報空間可視化部、視覚化された２次元の位置を指定することによって特定の文書を選択する文書選択部、および、選択された文書の内容を検索結果格納部より取り出して表示する文書内容表示部を設けたものである。

【００１２】また、請求項６に記載の発明に係る情報検索システムは、検索結果格納部内の文書の特定部分を切り出して格納する文書分類選択格納部を付加し、当該特定部分を入力テキストとして文書の自動分類および情報空間の可視化を行うものである。

【００１３】

【作用】請求項１に記載の発明における文書の自動分類方法は、文書の語句ベクトルとセルの語句ベクトルの距離が最小のセルをその文書の仮の所属セルとし、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近付け、またそのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて近付ける学習を行い、学習終了後、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断することにより、内容が類似した文書を一括して参照可能とする。

【００１４】また、請求項２に記載の発明における文書の自動分類方法は、上記学習に際して、語句の文書集合での出現割合によって近付ける割合を変化させることにより、文書のクラスタの階層性と文書クラスタの所属するセルの領域の包含関係が対応した文書のクラスタを作成する。

【００１５】また、請求項３に記載の発明における情報空間の可視化方法は、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句をそのセルを代表する語句としてその語句の文字列を

表示することにより、文書集合全体の概要が分かりやすい情報空間の可視化を可能とする。

【００１６】また、請求項４に記載の発明における情報空間の可視化方法は、セル間の境界線の属性を、隣接するセルに対応する語句ベクトルの距離に応じて変化させて表示することにより、内容の似たクラスタの領域が分かりやすい情報空間の可視化を可能とする。

【００１７】また、請求項５に記載の発明における情報検索システムは、情報検索によって得られた検索結果格納部内の文書集合を対象に、請求項１または２に記載された文書の自動分類方法を用いて文書の自動分類を行い、その自動分類された文書集合の情報空間を、請求項３または４に記載された情報空間の可視化方法によって視覚化し、視覚化された２次元の位置を指定することによって選択した文書の内容を、検索結果格納部より取り出して文書内容表示部に表示することにより、検索された文書集合の全体概要を見ながら個々の文書の内容を確かめることを可能にする。

【００１８】また、請求項６に記載の発明における情報検索システムは、検索結果格納部内の文書の特定部分を切り出して文書分類選択格納部に格納し、その特定部分を入力テキストとして文書の自動分類および情報空間の可視化を行うことにより、処理するデータ量を削減して処理時間を短縮する。

【００１９】

【実施例】

実施例１．以下、この発明の一実施例を図について説明する。図１はこの発明による文書の自動分類方法の一実施例における学習フェーズの処理の流れを示すフローチャートであり、図２は同じく分類フェーズの処理の流れを示すフローチャートである。このように、この実施例１による文書の自動分類方法は学習フェーズと分類フェーズとから成っており、以下、まず学習フェーズの動作について説明し、次に分類フェーズの動作について説明する。なお、この明細書中における「語句」という表現は、名詞、動詞などの通常の単語、および句や節など、テキストに含まれる意味のある文字列を表すものである。

【００２０】学習フェーズが開始されると、まずステップＳＴ１００において、文書ＤＯＣ－１，ＤＯＣ－２，＊・・・・，ＤＯＣ－ｉ，・・・・，ＤＯＣ－Ｎによって構成される文書集合ＤＯＣＵＭＥＮＴＳに含まれている異なった語句のリストを求めて、それを語句リストＷＯＲＤ－ＬＩＳＴとする。次にステップＳＴ１１０において、ノイズを削減するために、前記語句リストＷＯＲＤ－ＬＩＳＴの語句中より重要なもののみを選んで、それを語句リストＷＯＲＤ－ＬＩＳＴ２とする。例えば、出現頻度の高いものは一般的な語句であるため、分類という観点からは重要ではなく、また、出現頻度の低いものは特殊な語句であることが多く、これも分類という観点からは重要ではない。そこで、このステップＳＴ１１０では、語句リストＷＯＲＤ－ＬＩＳＴの各語句が文書集合ＤＯＣＵＭＥＮＴＳ中に含まれる頻度を数え、頻度が第１の閾値ＦＲＱＵＥＮＣＹ－ＬＯＷ以下の語句と頻度が第２の閾値ＦＲＱＵＥＮＣＹ－ＨＩＧＨ以上の語句を語句リストＷＯＲＤ－ＬＩＳＴの語句中より除き、それを語句リストＷＯＲＤ－ＬＩＳＴ２とする。なお、このようにして作成された語句リストＷＯＲＤ－ＬＩＳＴ２は、語句ｗｏｒｄ－１，ｗｏｒｄ－２，・・・・，ｗｏｒｄ－ｉ，・・・・，ｗｏｒｄ－ｎから構成されているものとする。

$$Vij = Fij \times \log(N/Nj) \quad \cdots\cdots \quad (1)$$

【００２３】ただし、上記（１）式において、Ｆｉｊは語句ｗｏｒｄ－ｊが文書ＤＯＣ－ｉに出現する頻度、Ｎは文書集合ＤＯＣＵＭＥＮＴＳの文書数、Ｎｊは語句ｗｏｒｄ－ｊを含む文書の数である。従って、語句ｗｏｒｄ－ｊが文書集合ＤＯＣＵＭＥＮＴＳのすべての文書に出現する場合は、Ｎｊ＝Ｎとなってｌｏｇ（Ｎ／Ｎｊ）＝０となるため、Ｖｉｊも０となる。これは分類という観点では、すべての文書に出現する語句は、その語句の有無によって文書を分けることができないため、その語句の重要度は０であることを表現している。



【００２１】次にステップＳＴ１２０において、２次元に配置されたセルの位置をＣＥＬＬ（ｘ，ｙ）とし、語句リストＷＯＲＤ－ＬＩＳＴ２を要素とする語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）を位置ＣＥＬＬ（ｘ，ｙ）のセルに対応するベクトルとする。なお、各語句ｗｏｒｄ－ｉの初期値は乱数などを使って任意の値にする。ただし、語句ベクトルは単位長に正規化する。次にステップＳＴ１３０に進み、文書集合ＤＯＣＵＭＥＮＴＳの各文書ＤＯＣ－ｉについて、語句リストＷＯＲＤ－ＬＩＳＴ２を要素とする語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉを作成する。語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉの各語句ｗｏｒｄ－ｊの値Ｖｉｊは、文書に出現する回数が多いほど　重要と考えられ、またその語句が出現する文書の数が少ないほど分類という観点からは重要であるので、そのような語句ほど値が大きくなるように、例えば次に示す（１）式によってその値を設定する。

【００２２】

【００２４】次にステップＳＴ１４０に進んで、後述するステップＳＴ１４１とステップＳＴ１４２の処理を、ｉ＝１からＮまで順にＴ回繰り返して実行する。なお、その場合、ｉはｉ＝Ｎの次はｉ＝１となるものとする。ステップＳＴ１４１では、各文書ＤＯＣ－ｉについて、その語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉと各位置ＣＥＬＬ（ｘ，ｙ）のセルの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）との距離を計算し、その距離が最小

のものをＣＥＬＬ（ｐ，ｑ）として、その位置のセルをその文書ＤＯＣ－ｉが所属する仮のセルとする。次にステップＳＴ１４２において、語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉをＶ、語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）をＷ（ｘ，ｙ）として、時刻ｔにおけるそ＊



＊のＷ（ｘ，ｙ）の値をＷ（ｘ，ｙ）（ｔ）とした時、そのＷ（ｘ，ｙ）の値を次の（２）式に従って更新する。

【００２５】

【数１】

$$W(x,y)(t+1) = \mathrm{Normalize}\left[W(x,y)(t) + \alpha(t) \times (V - W(x,y)(t))\right]$$

$$\alpha(t) = h(t) \times e^{\frac{(x-p)^2 + (y-q)^2}{\delta(t)^2}}$$

$$h(t) = H \times \frac{T-t}{T}$$

$$\delta(t) = \Delta \times \frac{T-t}{T} \qquad \cdots\cdots (2)$$

【００２６】ここで、上記（２）式において、Ｎｏｒｍａｌｉｚｅ（）はベクトルの長さを正規化する関数であり、ＨおよびΔは定数、α（ｔ）はＷ（ｘ，ｙ）をＶに近づける程度を表す学習係数である。この学習係数α（ｔ）は、時刻ｔが進むに従ってその大きさｈ（ｔ）と範囲δ（ｔ）が減少し、ｔ＝Ｔの時刻に０となる。

【００２７】ステップＳＴ１４０にて、このステップＳＴ１４１、ステップＳＴ１４２の処理がｉ＝１からＮまで順にＴ回繰り返されるとステップＳＴ１５０に進み、学習フェーズの一連の処理が終了する。

【００２８】このステップＳＴ１５０にて学習フェーズが終了すると、次に分類フェーズが開始される。この分類フェーズが開始されると、ステップＳＴ１６０においてまず、位置ＣＥＬＬ（ｘ，ｙ）のセルに属する文書の識別子ＩＤ－ｉを保存するためのリストをＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）として、そのリストＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）をｎｉｌに初期化する。次にステップＳＴ１７０に進んで、後述するステップＳＴ１７１の処理をｉ＝１からＮまで繰り返して実行する。このステップＳＴ１７１では、各文書ＤＯＣ－ｉについて、その語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉと各位置ＣＥＬＬ（ｘ，ｙ）のセルの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）との距離を計算し、それが最小であるセルの位置がＣＥＬＬ（ｐ，ｑ）であった場合に、リストＣＥＬＬ－Ｄｏｃ（ｐ，ｑ）にその文書ＤＯＣ－ｉの識別子ＩＤ－ｉを追加する。

【００２９】ステップＳＴ１７０にて、このステップＳＴ１７１の処理がｉ＝１からＮまで繰り返されるとステップＳＴ１８０に進み、この分類フェーズの一連の処理が終了する。なお、このようにして得られたリストＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）に属する文書が自動分類された文書クラスタである。

【００３０】なお、上記実施例１では、最初に与えられた文書集合に属する文書ＤＯＣ－ｉを自動的に分類するものについて説明したが、学習フェーズが終了した後、未知の文書についてもステップＳＴ１３０と同様の方法でその文書の語句ベクトルを作成し、ステップＳＴ１７１と同様の方法でその文書の属するセルを定めて分類に追加することにより、未知の文書を与えられた文書集合の自動分類と同一の基準で分類することが可能となる。

【００３１】さらに、この実施例１では、文書は１つのクラスタに分類されるものとして説明したが、図２のステップＳＴ１７１において、各位置ＣＥＬＬ（ｘ，ｙ）のセルの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）と各文書ＤＯＣ－ｉの語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉとの距離が一定の値以下の位置ＣＥＬＬ（ｘ，ｙ）のセルにすべての文書ＤＯＣ－ｉが所属するものとして、リストＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）にそれらの文書の識別子ＩＤ－ｉを追加することにより、文書が複数のクラスタに分類されるようにすることも可能である。

【００３２】また、図２のステップＳＴ１７１で、各位置ＣＥＬＬ（ｘ，ｙ）のセルの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）と各文書ＤＯＣ－ｉの語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉとの距離が小さいものから一定数の位置ＣＥＬＬ（ｘ，ｙ）のセルにすべての文書ＤＯＣ－ｉが所属するものとして、リストＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）にそれらの文書の識別子ＩＤ－ｉを追加することによっても、文書が複数のクラスタに分類されるようにすることが可能である。

【００３３】また、図２のステップＳＴ１７１で、各位置ＣＥＬＬ（ｘ，ｙ）のセルの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）と各文書ＤＯＣ－ｉの語句ベクトルＤＯＣ－Ｖｅｃｔｏｒ－ｉとの距離の分布を計算して、ローカルミニマムとなる位置ＣＥＬＬ（ｘ，ｙ）のセルに全ての文書ＤＯＣ－ｉが所属するものとして、リストＣＥＬＬ－Ｄｏｃ（ｘ，ｙ）にそれらの文書の識別子ＩＤ－ｉを追加することによっても、文書が複数のク

ラスタに分類されるようにすることが可能である。

【００３４】実施例２．実施例２はこの発明の文書の自動分類方法に関する他の実施例であり、上記実施例１では学習係数が語句の文書集合内での出現の仕方に関係なく一定であったのに対して、語句の文書集合内での出現の仕方に依存して学習係数を変化させている。なお、その学習係数の変化のさせ方については、例えば、ある語句ｗｏｒｄ－ｉが出現する文書の数をＮｉとするとき、学習フェーズの初期の段階ではＮｉが大きな語句の学習係数を、Ｎｉが小さな語句のそれよりも大きくしておき、学習が進むにつれてＮｉの小さな語句の方がＮｉの大きな語句よりも学習係数が大きくなるようにする。このように学習させることによって、Ｎｉの大きな一般的＊



＊な語句の要因を早く学習させることができるようになり、文書のクラスタの階層性と文書クラスタの属するセルの領域の包含関係が対応したものとなる。

【００３５】ここで、この実施例２の文書の自動分類方法における学習フェーズおよび分類フェーズでの処理の流れは、図１および図２のフローチャートに示した実施例１の場合と同様である。しかしながら、図１のステップＳＴ１４２にて用いられているＷ（ｘ，ｙ）（ｔ）の値を更新するための式として、例えば次に示す（３）式を用いている点で異なっている。

【００３６】

【数２】

$$W(x,y)(t+1) = \mathrm{Normalize}\left[W(x,y)(t) + \alpha(t,Ni) \times (V - W(x,y)(t))\right]$$

$$\alpha(t,Ni) = h(t,Ni) \times e^{-\frac{(x-p)^2+(y-q)^2}{\delta(t,Ni)^2}}$$

$$h(t,Ni) = H(Ni) \times \frac{Th(Ni)-t}{Th(Ni)}$$

$$\delta(t,Ni) = \Delta(Ni) \times \frac{Ts(Ni)-t}{Ts(Ni)}$$

$$H(Ni) = H0 \times [1-G(Ni)]$$

$$\Delta(Ni) = \Delta 0 \times [1-G(Ni)]$$

$$Th(Ni) = Th0 \times G(Ni)$$

$$Ts(Ni) = Ts0 \times G(Ni)$$

$$G(Ni) = \frac{\log(N/Ni)}{\log N} \qquad \cdots\cdots (3)$$

【００３７】なお、上記（３）式において、Ｈ０，Δ０，Ｔｈ０，Ｔｓ０は定数であり、Ｎｉは１からＮまでの整数値をとる。ここで、関数Ｇ（ｘ）がＧ（１）＝１、Ｇ（Ｎ）＝０となる単調減少関数であれば、上記 ※

※（３）式におけるＧ（Ｎｉ）は次の（４）式に示すような他の関数であってもよい。

【００３８】

$$G(Ni) = (N-Ni) / (N-1) \qquad \cdots\cdots (4)$$

【００３９】実施例３．図３はこの発明による情報空間の可視化方法の一実施例における処理の流れを示すフローチャートであり、図４は２次元に配置されたセルの配置例を示す説明図、図５は可視化された情報空間の表示例を示す説明図である。この図４および図５において、５は２次元に配置されたセルであり、図４においてはその各々が配置されている位置がＣＥＬＬ（０，０），ＣＥＬＬ（０，１），・・・・，ＣＥＬＬ（３，３）で表されている。また、図５において、６はセル５を代表する語句としてそのセル５内に表示された意味のある文字列であり、隣り合ったセル５の間で代表する語句が同一である場合にはその境界線を消去し、それに１つの文字列６を表示している。なお、図４では各セル５が６角形であるものを示したが、４角形など他の形状であってもさしつかえない。

【００４０】次にその動作を図３のフローチャートに従って説明する。まず、図１に示した実施例１あるいは実施例２の学習フェーズが終了した後、各位置ＣＥＬＬ（ｘ，ｙ）のセル５の語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の語句をその値の順にソートする。次にステップＳＴ２１０に進み、その値の大きい順に、あらかじめ定められた数の語句を選択して、それをその位置

ＣＥＬＬ（ｘ，ｙ）のセル５を代表する語句とする。次にステップＳＴ２２０において、その選択された語句の文字列６をそれぞれの位置ＣＥＬＬ（ｘ，ｙ）のセル５に表示する。

【００４１】以下、この文字列６の表示を図５に従って具体的に説明する。図５は国際特許分類のサブクラスＧ０６Ｆに分類されている特許文書に関して、同一出願人の特許文書４１件について自動分類し、その情報空間を可視化した場合の表示例を示したものであり、各位置ＣＥＬＬ（ｘ，ｙ）のセル５の語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値が最大の語句を１つだけ選択し、その語句の文字列６を各セル５に表示したものである。なお、この図５においては、表示を見やすくするため、隣接するセル５の相互で代表する語句が同一である場合には、その境界線を消すとともに、その中に文字列６を１つだけ表示するようにしている。例えば、図５の右下のセル５とその左隣のセル５とは代表する語句が同一であるため、両者の間の境界線が消去され、その一方（右下隅のセル５）にのみ共通の文字列６として「処理装置」が表示されている。

【００４２】また、この図５では、その右上の部分に「ＣＰＵ」、「プロセッサ」、「プログラム」などの関係の深い語句の文字列６が表示されたセル５が配置されており、左上の部分には「ディスク装置」、「記憶装置」という関係の深い語句の文字列６が表示されたセル５が配置されている。さらに、その下側には「電力系統」と「知識ベース」の文字列６が表示されたセル５が隣接して配置されているが、これは電力系統の監視に知識ベースを持つエキスパートシステムが利用されていることが推測できる。このように、この実施例３の情報空間の可視化方法によれば、それぞれの代表的な語句の関連が深いセル５が互いに近くになるように配置されて可視化されることとなる。

【００４３】なお、この実施例３では、それぞれの位置ＣＥＬＬ（ｘ，ｙ）の語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値が最大の語句を１つ選んで、その文字列６を該当するセル５に表示する場合について説明したが、語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値の大きいものから順に一定個数の語句を選択して、その文字列６を表示するようにしても、また、語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値が一定値以上のものをすべて表示するようにしてもよい。なお、その場合、語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値に応じて語句の重要度が区別できるように、文字列６の大きさや書体、さらには表示色などの文字属性を変えるようにしてもよい。

【００４４】さらに、各語句ｗｏｒｄ－ｊの各位置ＣＥＬＬ（ｘ，ｙ）のセルでの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の値をグラフとして表示するようにしてもよく、また語句ｗｏｒｄ－ｊの各位置ＣＥＬＬ（ｘ，ｙ）のセルでの値の分布を計算して、ローカルマキシマムとなるセル５の位置ＣＥＬＬ（ｘ，ｙ）にその語句ｗｏｒｄ－ｊを表示するようにしてもよい。

【００４５】実施例４．図６はこの発明による情報空間の可視化方法の他の実施例における処理の流れを示すフローチャートであり、図７は可視化された情報空間の表示例を示す説明図である。図７において、５はセルであり、７はそのセル５を代表する語句である。また、８は隣接するセル５間の境界線で、セル５を代表する語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｘ，ｙ）の距離により、その属性が変えられて表示されるものである。

【００４６】次にその動作を図６のフローチャートに従って説明する。まずステップＳＴ３００において、互いに隣接したセル５をそれぞれセルａ、セルｂとしたとき、それらの境界をＥＤＧＥ（ａ，ｂ）とする。次にステップＳＴ３１０で、すべての境界ＥＤＧＥ（ａ，ｂ）について、セルａの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ａｘ，ａｙ）とセルｂの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｂｘ，ｂｙ）の距離を計算する。次にステップＳＴ３２０に進んで、ステップＳＴ３１０で算出された各境界ＥＤＧＥ（ａ，ｂ）における距離の値を、その最大値のものが１となるように正規化する。次にステップＳＴ３３０で、各境界ＥＤＧＥ（ａ，ｂ）を示す境界線８の属性値を、その境界ＥＤＧＥ（ａ，ｂ）の距離の値に従って、あらかじめ定めておいた種類や太さなどを表すものに割り当てる。次にステップＳＴ３４０において、セルａとセルｂの境界ＥＤＧＥ（ａ，ｂ）の境界線８を、その割り当てられた属性によって表示し、ステップＳＴ３５０にて一連の処理を終了する。

【００４７】以下、この境界線８の表示を図７を用いて具体的に説明する。ここでは説明を簡単化するため、境界線８の属性の種類は太線と破線の２種類とし、太線は隣り合うセルａの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ａｘ，ａｙ）とセルｂの語句ベクトルＣＥＬＬ－Ｖｅｃｔｏｒ（ｂｘ，ｂｙ）の距離が大きく、破線はその距離が小さいことを表すものとする。ここで、位置ＣＥＬＬ（ｘ，ｙ）のセル５における代表的な語句７をＷＯＲＤｘｙとすると、図７は次のことを表していると解釈できる。まず、可視化された情報空間が大きく分けて３つの領域に別れている。すなわち、第１の領域は位置ＣＥＬＬ（０，２）、ＣＥＬＬ（１，２）、ＣＥＬＬ（０，３）およびＣＥＬＬ（１，３）の４つのセル５による領域である。第２の領域は位置ＣＥＬＬ（０，０）、ＣＥＬＬ（１，０）、ＣＥＬＬ（２，０）、ＣＥＬＬ（３，０）、ＣＥＬＬ（０，１）、ＣＥＬＬ（１，１）、ＣＥＬＬ（２，１）、ＣＥＬＬ（２，２）、ＣＥＬＬ（３，２）、ＣＥＬＬ（２，３）およびＣＥＬＬ（３，３）の１１個のセル５による領域である。第３の領域は位置ＣＥＬＬ（３，１）の１つのセル５による領域である。

【００４８】また、第１の領域の各セル５を代表する語

句７であるＷＯＲＤ０２、ＷＯＲＤ１２、ＷＯＲＤ０３およびＷＯＲＤ１３は互いに連想関係にあり、それぞれのセル５に対応する文書も内容が近い。一方、ＷＯＲＤ０２とＷＯＲＤ０１で代表されるセル５、ＷＯＲＤ１２とＷＯＲＤ０１で代表されるセル５、ＷＯＲＤ１２とＷＯＲＤ１１で代表されるセル５、ＷＯＲＤ１２とＷＯＲＤ２２で代表されるセル５、ＷＯＲＤ１３とＷＯＲＤ２２で代表されるセル５、ＷＯＲＤ１３とＷＯＲＤ２３で代表されるセル５は互いに隣接していても、対応する文書は近い関係にはない。さらに第２の領域内においても、ＷＯＲＤ０１とＷＯＲＤ１０で代表されるセル５、およびＷＯＲＤ１１とＷＯＲＤ１０で代表されるセル５は近い関係にあるが、ＷＯＲＤ０１とＷＯＲＤ１１で代表されるセル５は隣接していても近い関係にはない。

【００４９】実施例５．図８はこの発明による情報検索システムの一実施例の機能構成を示す構成図である。図において、１は文書データベース、２は情報検索部、３は検索条件入力部、４は検索結果表示部であり、これらは図１３に同一符号を付した従来のそれらと同一、もしくは相当部分であるためその説明を省略する。

【００５０】また、９は情報検索部２によって検索された文書集合を格納するための検索結果格納部であり、１０はこの検索結果格納部９に格納された文書集合を対象にして、請求項１または２に記載された文書の自動分類方法に従って文書の自動分類を行う自動分類部、１１は請求項３または４に記載された情報空間の可視化方法に従って、この自動分類部１０で自動分類された文書クラスタの代表する語句を２次元に視覚化する情報空間可視化部である。１２はこの情報空間可視化部１１によって視覚化された２次元の位置を指定することによって、自動分類された文書の中から特定の文書の選択を行う文書選択部であり、１３はこの文書選択部１２によって選択された文書の内容を検索結果格納部９より取り出して表示する文書内容表示部である。

【００５１】次に動作について説明する。ここで、図９はこの実施例５による情報検索システムの処理の流れを示すフローチャートである。まず、ステップＳＴ４００において、検索条件入力部３より検索条件を入力する。この検索条件はキーワードの論理積や論理和によるものである。次にステップＳＴ４１０において、情報検索部２が文書データベース１を検索してその検索条件に合う文書を抽出し、ステップＳＴ４２０でその検索結果を検索結果表示部４に表示する。なお、この検索結果は通常は検索条件に該当する文書の数であり、必要に応じて文書のタイトルや概要なども表示することがある。次にステップＳＴ４３０に進み、情報検索部２で検索された文書が、内容および数の観点から見て、検索の初期の目的を満たしているか否かを利用者が判断する。その結果、初期の目的を満たしていなければステップＳＴ４００に戻り、新たな検索条件で再検索を行う。

【００５２】一方、初期の目的を満たしている場合には、ステップＳＴ４４０にて検索結果の文書集合を検索結果格納部９に格納する。次にステップＳＴ４５０に進み、検索結果格納部９に格納されている文書集合を対象に、自動分類部１０で文書の自動分類を行う。なお、この文書の自動分類は実施例１もしくは実施例２で説明した文書の自動分類方法によって実現される。次にステップＳＴ４６０で情報空間可視化部１１によって、自動分類部１０が前述のようにして自動分類した情報空間を、代表的な語句で関連の深いものが近くにくるように配置されたキーワードマップの形に可視化表示する。なお、この情報空間の可視化も実施例３もしくは実施例４で説明した情報空間の可視化方法によって実現される。

【００５３】次にステップＳＴ４７０において、利用者がこの情報空間可視化部１１によって可視化されたキーワードマップを参照して、文書選択部１２にて関心のあるセルを選択する。セルが選択されると処理はステップＳＴ４８０に進み、選択されたセルに対応する文書集合がタイトルリストの形で表示される。次にステップＳＴ４９０において、利用者がこのタイトルリストの形で表示された文書集合を参照し、文書選択部１２にて関心のある文書を１つ選択する。文書が選択されると処理はステップＳＴ５００に進み、文書内容表示部１３は文書結果格納部９より選択された文書の内容を取り出して表示する。次にステップＳＴ５１０において、利用者が文書内容表示部１３に表示された文書の内容を参照し、満足するものであるか否かを判断する。その結果、満足できるものであった場合にはステップＳＴ５２０に進み、一連の処理を終了する。

【００５４】一方、満足できるものではなかった場合には、ステップＳＴ４９０に戻って表示されているタイトルリストの中から別の文書を指定してその内容を参照したり、ステップＳＴ４７０に戻ってキーワードマップの別のセルを選択する。このようにして、利用者はキーワードマップで可視化された情報空間を見ながら、満足するまで検索を繰り返す。なお、文書データベース１の大きさが小さい場合には、検索結果格納部９は文書データベース１で代用することも可能である。また、ステップＳＴ５２０での終了は、選択された検索結果の文書集合を対象とした自動分類、情報空間可視化に対するものであり、満足するものがなかった場合にはステップＳＴ４００に戻り、新たな検索条件を検索条件入力部３に入力する。

【００５５】次に、ステップＳＴ４６０からＳＴ５００までの処理を図について詳細に説明する。図１０はこの実施例５による情報検索システムの実行時のスナップショットを示す説明図であり、図中、１４は情報空間可視化部１１で可視化されたキーワードマップがステップＳＴ４６０において表示されるウィンドウ、１５は選択されたセルに対応する文書のタイトルリストがステップＳ

Ｔ４８０において表示されるウィンドウ、１６は選択された文書の内容がステップＳＴ５００において表示されるウィンドウである。

【００５６】ウィンドウ１４に表示されたキーワードマップの中の「レジスタ」と表示されているセルを、利用者がマウスなどでポインティングすることによって選択すると、そのセルに対応する自動分類された文書クラスタの文書のタイトルリストがウィンドウ１５に表示される。この例では、「マイクロコンピュータ」というタイトルの文書と、「データ処理回路」というタイトルの文書がクラスタになっていたことがわかる。次に、このウィンドウ１５上で利用者が「マイクロコンピュータ」の文書を、マウスなどでポインティングすることによって選択すると、その文書の内容がウィンドウ１６に表示される。利用者はこのウィンドウ１６の表示を参照して、それが満足するものであるか否かを判断する。

【００５７】なお、この実施例５では、図９のステップＳＴ４７０で利用者が関心のあるセルを選択して、ステップＳＴ４８０で選択されたセルに対応する文書集合をタイトルリストの形で表示した後、ステップＳＴ４９０で利用者がそのタイトルリストを見て関心のある文書を１つ選択し、ステップＳＴ５００でその内容を表示するものについて説明したが、ステップＳＴ４７０で利用者が選択したセルに対応する文書集合に含まれている文書の数が１個、または表示画面の制約から決まる所定の個数よりも小さい場合には、ステップＳＴ４８０およびＳＴ４９０を省略して文書の内容を表示するようにしてもよい。

【００５８】また、上記実施例５では、文書データベース１の規模が大きく、前処理として検索条件入力部３より入力したキーワードによる検索条件により情報検索を行って、自動分類や情報空間の可視化の対象となる文書の数を絞り込んだ場合について示したが、文書データベース１の規模が小さい場合には、文書データベース１の内容をすべて検索結果格納部９に入れておき、ステップＳＴ４４０からスタートするようにしてもよい。これはステップＳＴ４００で文書データベース１のすべての文書が該当する検索条件を入力し、ステップＳＴ４３０で「ＹＥＳ」と判断したことに対応する。

【００５９】さらに、検索条件入力部３に入力するキーワードの候補となるものを、文書内容表示部１３に表示されたテキストの文字列からあらかじめ抽出しておいて、それを選択することにより簡易に入力できるようにしてもよい。

【００６０】また、ステップＳＴ４７０で利用者が関心のあるセルを１つまたは複数個選択して、それらのセルに対応する文書集合を検索結果格納部９に格納して、ステップＳＴ４５０に移れるようにしてもよい。

【００６１】また、ステップＳＴ５００で文書の内容を表示するとき、情報空間可視化部１１で可視化されたキーワードマップのセルを代表する語句の文字列を、表示色などの属性を変えて分かりやすく表示するようにしてもよい。

【００６２】また、検索結果格納部９に格納された文書集合のデータを文書データベース１のデータとして切り替えられるようにしてもよい。

【００６３】実施例６．図１１はこの発明による情報検索システムの他の実施例の機能構成を示す構成図で、相当部分には図８と同一符号を付してその説明を省略する。図において、１７は情報検索部２にて検索され、検索結果格納部９に格納された文書集合の各文書の特定部分を切り出して格納する文書分類選択格納部である。なお、自動分類部１０はこの文書分類選択格納部１７に格納されている文書の特定部分を入力テキストとして文書の自動分類を行うものであり、情報空間可視化部１１は当該文書の特定部分を入力テキストとして自動分類された情報空間の可視化を行うものである。

【００６４】次に動作について説明する。ここで、図１２はこの実施例６による情報検索システムの処理の流れを示すフローチャートである。まず、ステップＳＴ４００からステップＳＴ４４０において、図９に同一のステップ番号を付した実施例５で説明したのと同様の処理が実行される。その後ステップＳＴ４４１に進み、検索結果格納部９に格納されている文書から、あらかじめ定められた特定部分を選択して切り出し、それを文書分類選択格納部１７に格納する。

【００６５】この特定部分の選択の方法としては、文書の種類に応じて、例えば文書の概要や前書きの第１段落などを選択する。また文書記述のためのＩＳＯ（International Organization for Standardization：国際標準化機構）標準である、ＳＧＭＬ（Standard Generalized Markup Language）などの規格に準拠して作成されたタグ付きの文書では、文書の連続した部分だけではなく、文書の連続していない複数の場所から選択することを自動的に行うこともできる。

【００６６】次にステップＳＴ４５１において、自動分類部１０がこの文書分類選択格納部１７に格納されている文書集合を対象に、実施例１もしくは実施例２で説明した文書の自動分類方法による文書の分類が行われる。以下、ステップＳＴ４６０からステップＳＴ５２０において、図９に同一のステップ番号を付した実施例５で説明したのと同様の処理が実行される。このように、この実施例６では、検索結果格納部９の内容よりデータ量がはるかに少ない文書分類選択格納部１７の内容を用いて、文書の自動分類および情報空間の可視化が行われることになる。

【００６７】なお、上記各実施例では、独立した文書を対象とするものを示したが、文書が互いにリンクで結ば

れたハイパーテキストを対象としてもよく、上記実施例と同様の効果を奏する。

【００６８】また、上記各実施例では、１箇所のコンピュータにデータベースとして蓄えられている文書を対象としたものについて説明したが、コンピュータネットワークによって接続された複数のコンピュータに分散して蓄えられた文書を対象にしてもよく、上記実施例と同様の効果を奏する。

【００６９】

【発明の効果】以上のように、請求項１に記載の発明によれば、文書の語句ベクトルとセルの語句ベクトルの距離が最小のセルをその文書の仮の所属セルとして、その所属セルの語句ベクトルの値をその文書の語句ベクトルの値に近付けるとともに、そのセル近傍のセルの語句ベクトルの値もその文書の語句ベクトルへの近さの度合いに応じて近付ける学習を行って、その学習の終了後に、各セルの語句ベクトルと文書の語句ベクトルの距離が最小のセルをその文書の所属セルとして、同じセルに所属する文書を内容が類似した文書のクラスタと判断するように構成したので、文書を語句のベクトルとして表現し、そのベクトル表現された文書を自動分類することが可能となり、内容の類似した文書を一括して見ることができる文書の自動分類方法が得られる効果がある。

【００７０】また、請求項２に記載の発明によれば、学習の際に語句の文書集合における出現割合に基づいて近付ける割合を変化させるように構成したので、語句の文書集合での出現の分布に依存した学習を行わせることが可能となり、文書のクラスタの階層性と文書クラスタの属するセルの領域の包含関係が対応した文書のクラスタを作成できる効果がある。

【００７１】また、請求項３に記載の発明によれば、自動分類された情報空間について、セルに対応する語句ベクトルの値が大きなものから一定数、もしくは所定の閾値以上の語句をそのセルを代表する語句としてその文字列を表示するように構成したので、自動分類された文書集合を代表する語句を関連の深いセルが近くに配置されるように表示して、文書集合の情報空間を可視化することが可能となって、文書集合全体の概要が分かりやすい情報空間の可視化方法を得ることができ、さらに、どのようなキーワードがあるかを容易に知ることが可能となるばかりか、分類されたクラスタの相互関係が把握しやすくなる効果がある。

【００７２】また、請求項４に記載の発明によれば、隣接するセルに対応する語句ベクトルの距離に応じて、セル間の境界線の属性を変化させて表示するように構成したので、隣り合ったセルの類似度を判断することが容易となって、文書集合全体の概要がより分かりやすいものとなり、また内容の似たクラスタの領域が分かりやすくなる効果がある。

【００７３】また、請求項５に記載の発明によれば、情報検索によって検索結果格納部に格納された文書集合を対象に、請求項１または２に記載した文書の自動分類方法による文書の自動分類を行い、自動分類された文書集合の情報空間を、請求項３または４に記載した情報空間の可視化方法で視覚化し、視覚化された２次元の位置指定によって選択した文書の内容を検索結果格納部より取り出して表示するように構成したので、可視化された情報空間のセルを指定してそのセルに属する文書のリストを表示し、さらに表示されたリストの文書を指定することによって所望の文書の内容を表示することが可能となり、検索された文書集合の全体概要を見ながら個々の文書の内容を確かめることができる情報検索システムが得られる効果がある。

【００７４】また、請求項６に記載の発明によれば、検索結果格納部内の文書の特定部分を切り出して文書分類選択格納部に格納しておき、文書の自動分類および情報空間の可視化をその文書の特定部分を入力テキストとして行うように構成したので、検索結果格納部よりはるかにデータ量の少ない文書部分選択格納部の内容を用いて、文書の自動分類および情報空間の可視化を行うことが可能となり、処理するデータ量が減ることによって処理が高速化される効果がある。

【図面の簡単な説明】

【図１】　この発明の実施例１による文書の自動分類方法の学習フェーズの処理の流れを示すフローチャートである。

【図２】　上記実施例における分類フェーズの処理の流れを示すフローチャートである。

【図３】　この発明の実施例３による情報空間の可視化方法の処理の流れを示すフローチャートである。

【図４】　上記実施例におけるセルの配置例を示す説明図である。

【図５】　上記実施例における可視化された情報空間の表示例を示す説明図である。

【図６】　この発明の実施例４による情報空間の可視化方法の処理の流れを示すフローチャートである。

【図７】　上記実施例における可視化された情報空間の表示例を示す説明図である。

【図８】　この発明の実施例５による情報検索システムの機能構成を示す構成図である。

【図９】　上記実施例の処理の流れを示すフローチャートである。

【図１０】　上記実施例における実行時のスナップショットを示す説明図である。

【図１１】　この発明の実施例６による情報検索システムの機能構成を示す構成図である。

【図１２】　上記実施例の処理の流れを示すフローチャートである。

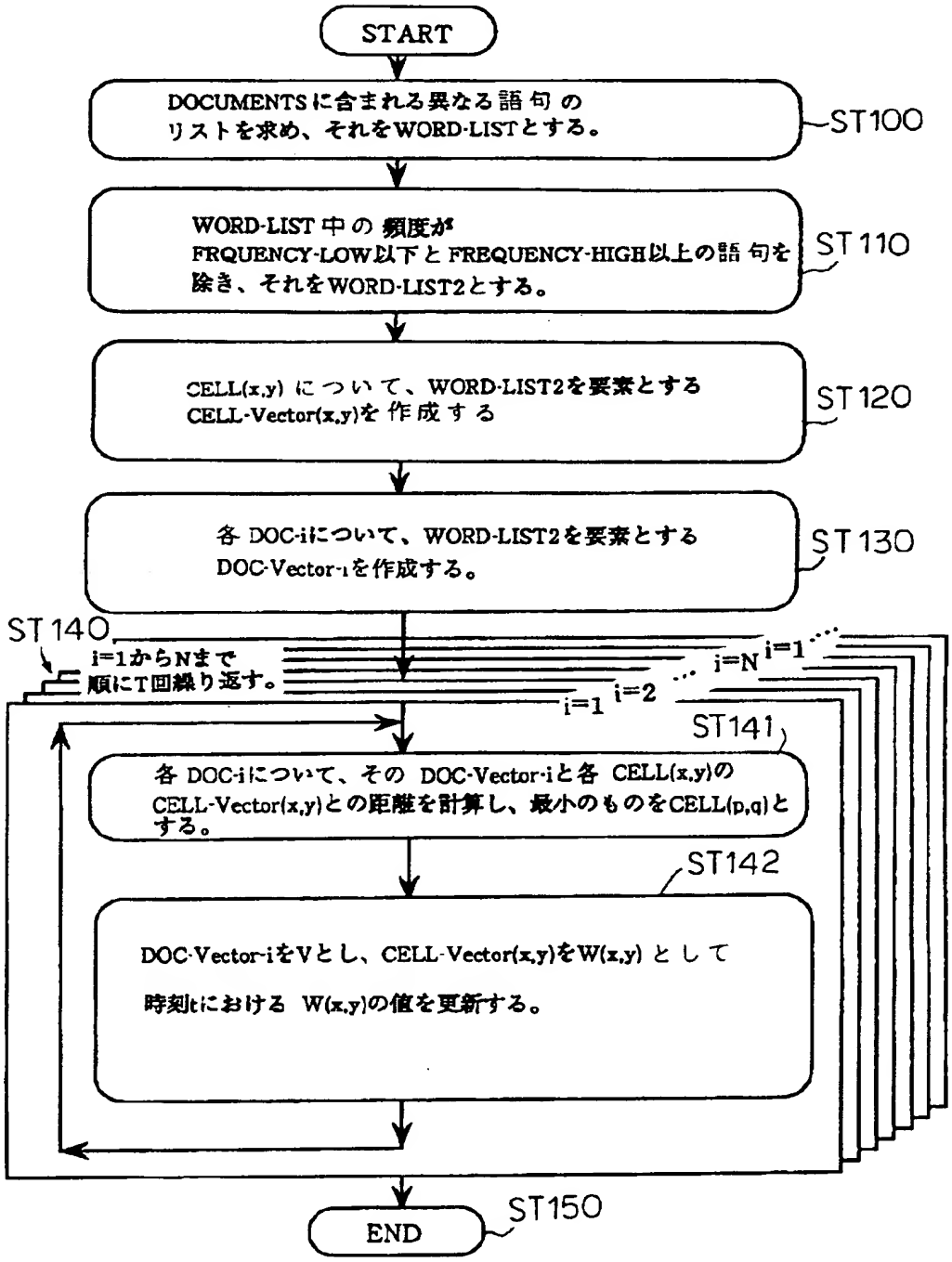
【図１３】　従来の情報検索システムの機能構成を示す構成図である。

19　　20

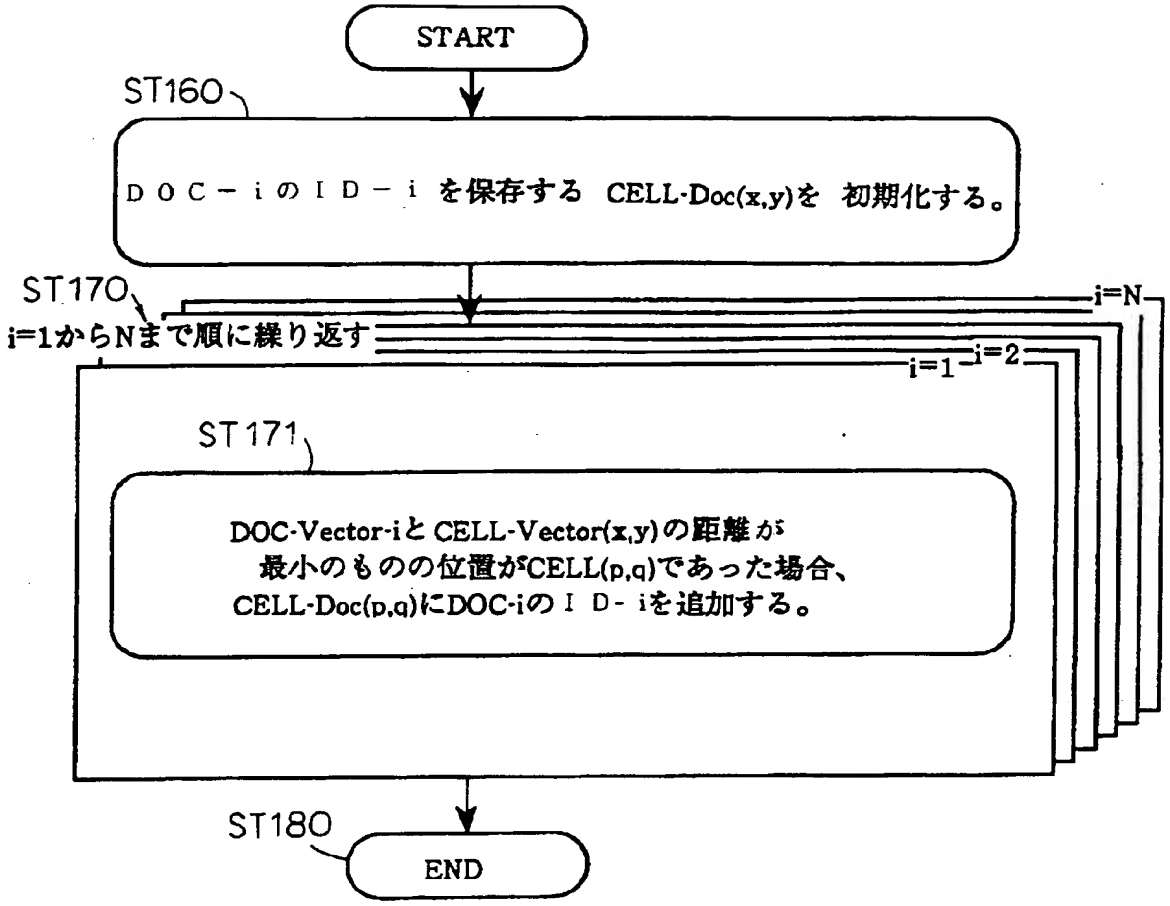
【符号の説明】

１　文書データベース、２　情報検索部、３　検索条件入力部、４　検索結果表示部、５　セル、６　文字列、＊８　境界線、９　検索結果格納部、１０　自動分類部、１１　情報空間可視化部、１２　文書選択部、１３　文書内容表示部、１７　文書分類選択格納部。

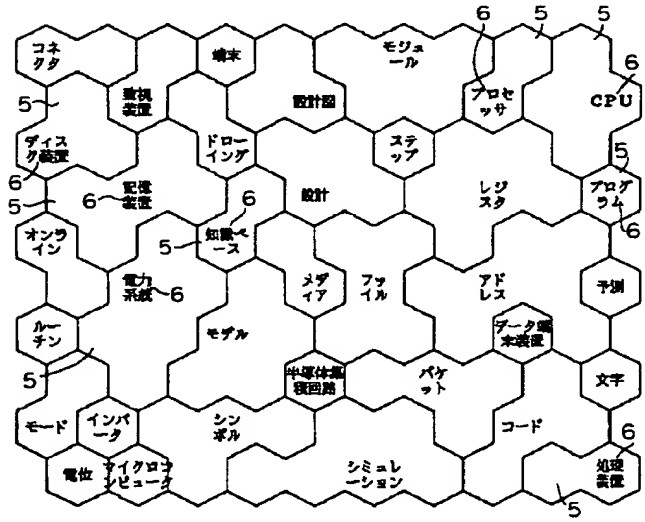
【図１】



【図２】

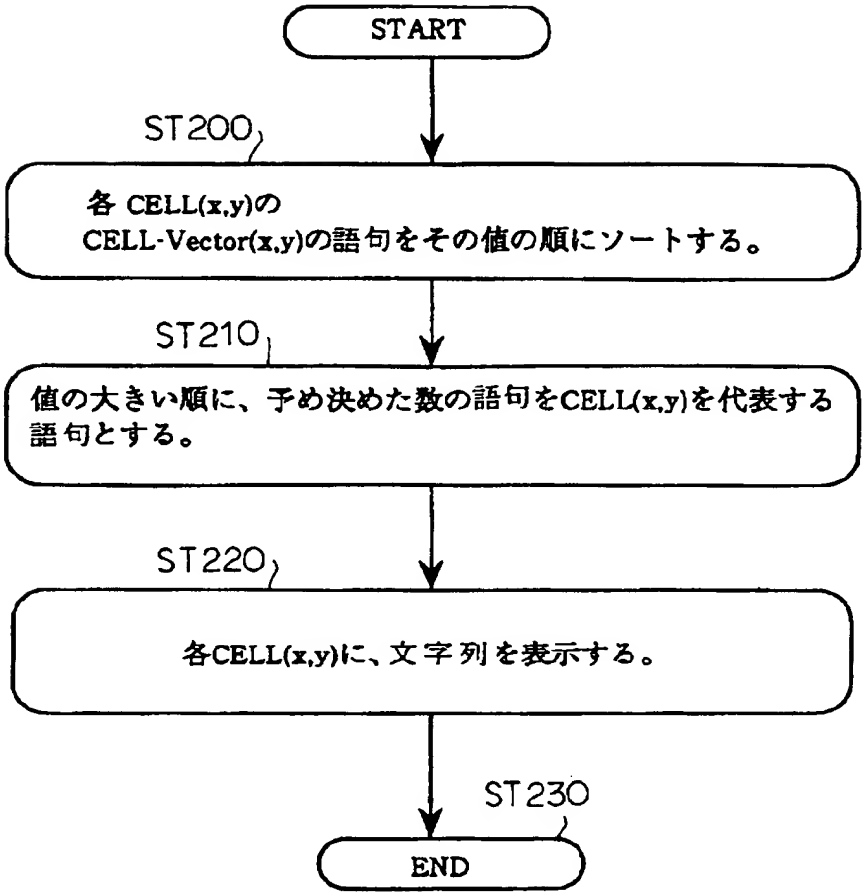

START
ST160
ＤＯＣ－ｉのＩＤ－ｉを保存する CELL-Doc(x,y)を 初期化する。
ST170
i=1からNまで順に繰り返す
i=N
i=2
i=1
ST171
DOC-Vector-iとCELL-Vector(x,y)の距離が
最小のものの位置がCELL(p,q)であった場合、
CELL-Doc(p,q)にDOC-iのＩＤ- iを追加する。
ST180
END

【図５】

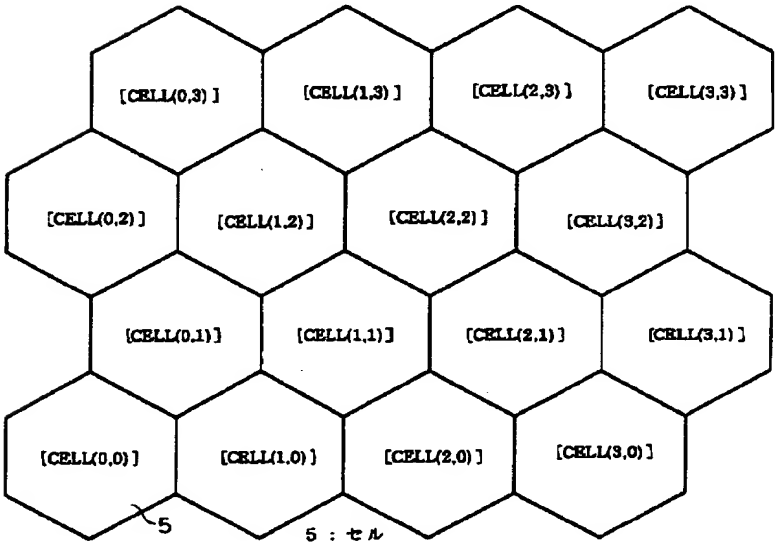

コネクタ
端末
モジュール
プロセッサ
CPU
表示装置
設計図
ディスク装置
ドローイング
ステップ
記憶装置
設計
レジスタ
プログラム
オンライン
知識ベース
電力系統
メディア
ファイル
アドレス
予測
ルーチン
モデル
データ端末装置
半導体集積回路
パケット
文字
モード
インバータ
シンボル
コード
電位
マイクロコンピュータ
シミュレーション
処理装置
6
5

6：文字列

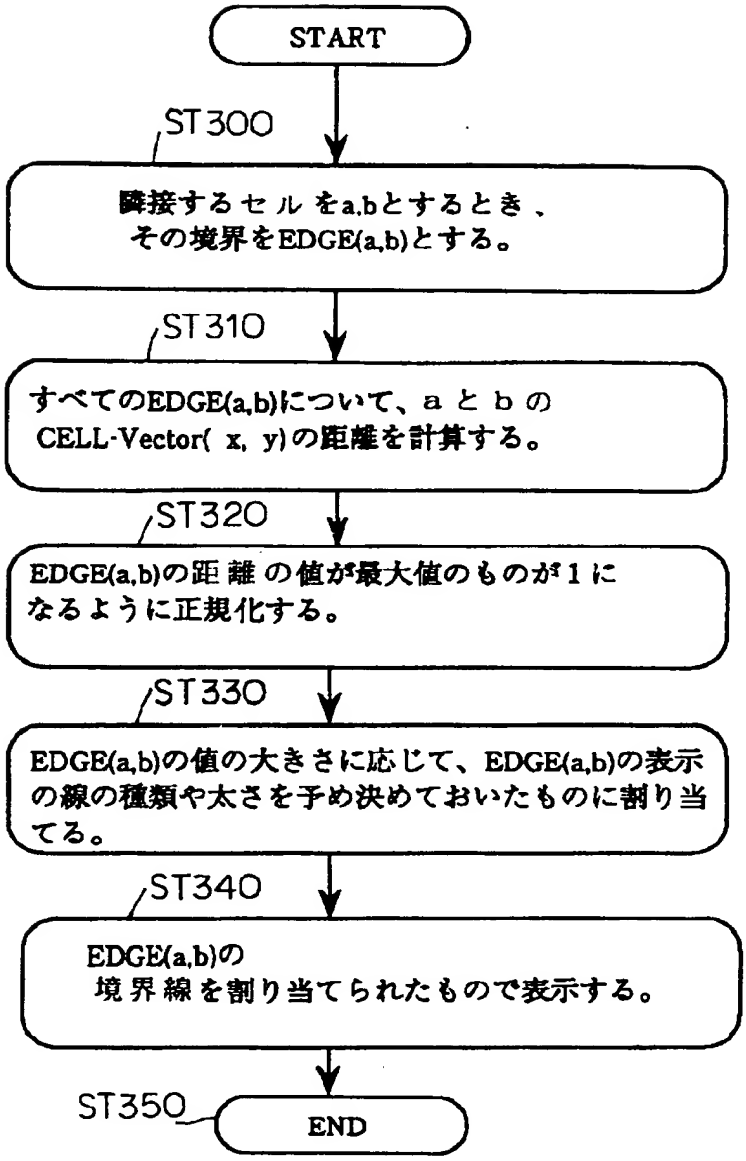
【図３】


START
ST200
各 CELL(x,y)の
CELL-Vector(x,y)の語句をその値の順にソートする。
ST210
値の大きい順に、予め決めた数の語句をCELL(x,y)を代表する語句とする。
ST220
各CELL(x,y)に、文字列を表示する。
ST230
END

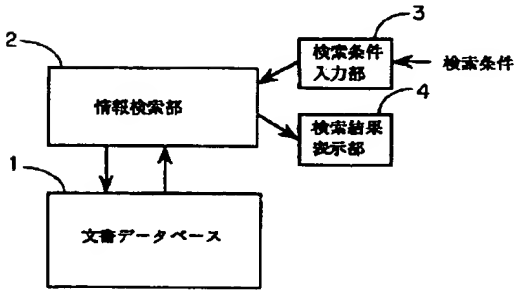
【図４】


[CELL(0,3)]
[CELL(1,3)]
[CELL(2,3)]
[CELL(3,3)]
[CELL(0,2)]
[CELL(1,2)]
[CELL(2,2)]
[CELL(3,2)]
[CELL(0,1)]
[CELL(1,1)]
[CELL(2,1)]
[CELL(3,1)]
[CELL(0,0)]
[CELL(1,0)]
[CELL(2,0)]
[CELL(3,0)]
5
5 : セル

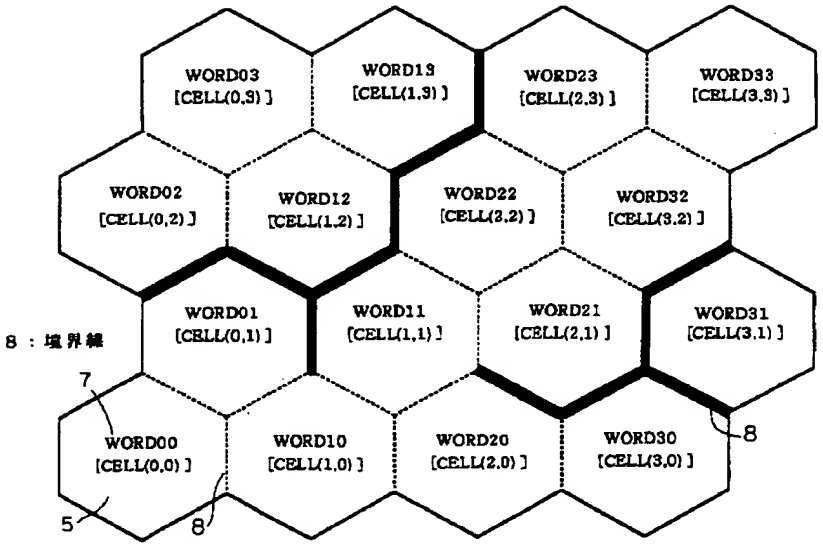
【図６】


START
ST300
隣接するセルをa,bとするとき、
その境界をEDGE(a,b)とする。
ST310
すべてのEDGE(a,b)について、ａとｂの
CELL-Vector( x, y)の距離を計算する。
ST320
EDGE(a,b)の距離の値が最大値のものが１に
なるように正規化する。
ST330
EDGE(a,b)の値の大きさに応じて、EDGE(a,b)の表示
の線の種類や太さを予め決めておいたものに割り当
てる。
ST340
EDGE(a,b)の
境界線を割り当てられたもので表示する。
ST350
END

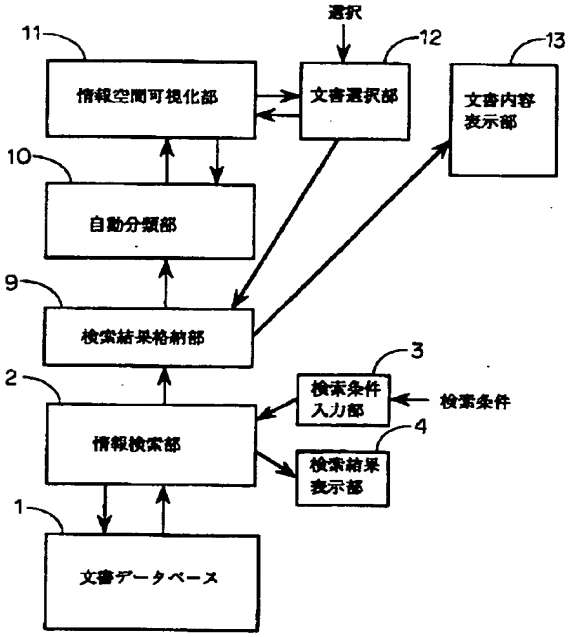
【図１３】


2
3
検索条件
入力部
検索条件
情報検索部
4
検索結果
表示部
1
文書データベース

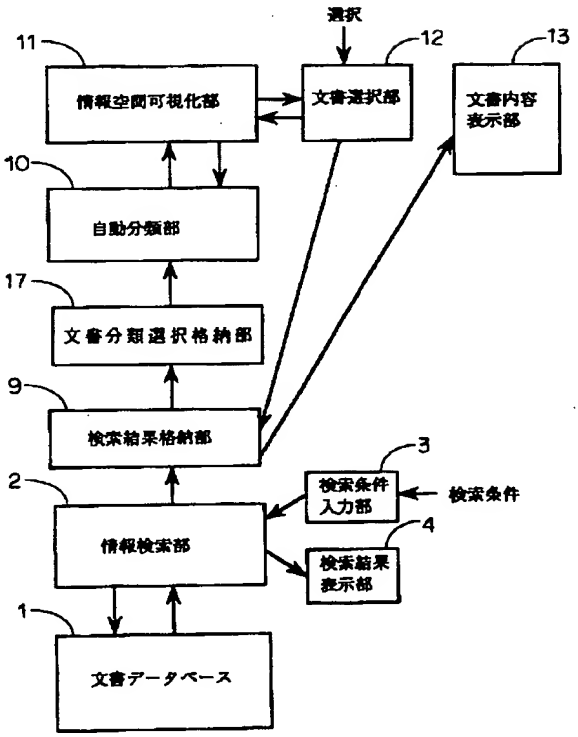
【図７】


WORD03
[CELL(0,3)]
WORD13
[CELL(1,3)]
WORD23
[CELL(2,3)]
WORD33
[CELL(3,3)]
WORD02
[CELL(0,2)]
WORD12
[CELL(1,2)]
WORD22
[CELL(2,2)]
WORD32
[CELL(3,2)]
WORD01
[CELL(0,1)]
WORD11
[CELL(1,1)]
WORD21
[CELL(2,1)]
WORD31
[CELL(3,1)]
WORD00
[CELL(0,0)]
WORD10
[CELL(1,0)]
WORD20
[CELL(2,0)]
WORD30
[CELL(3,0)]
8：境界線
7
8
5
8

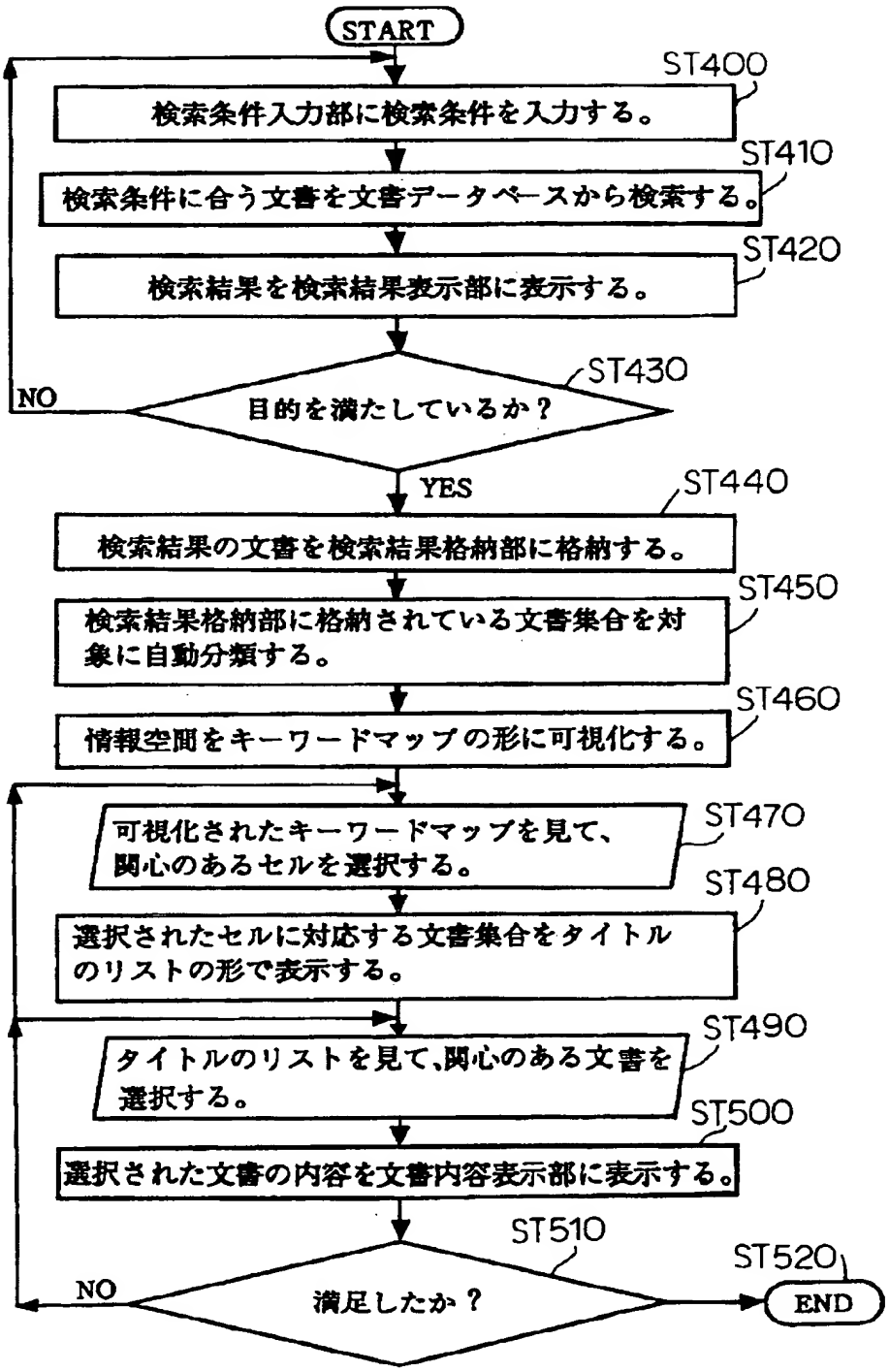
【図８】


選択
11
12
13
情報空間可視化部
文書選択部
文書内容
表示部
10
自動分類部
9
検索結果格納部
3
検索条件
入力部
検索条件
2
情報検索部
4
検索結果
表示部
1
文書データベース

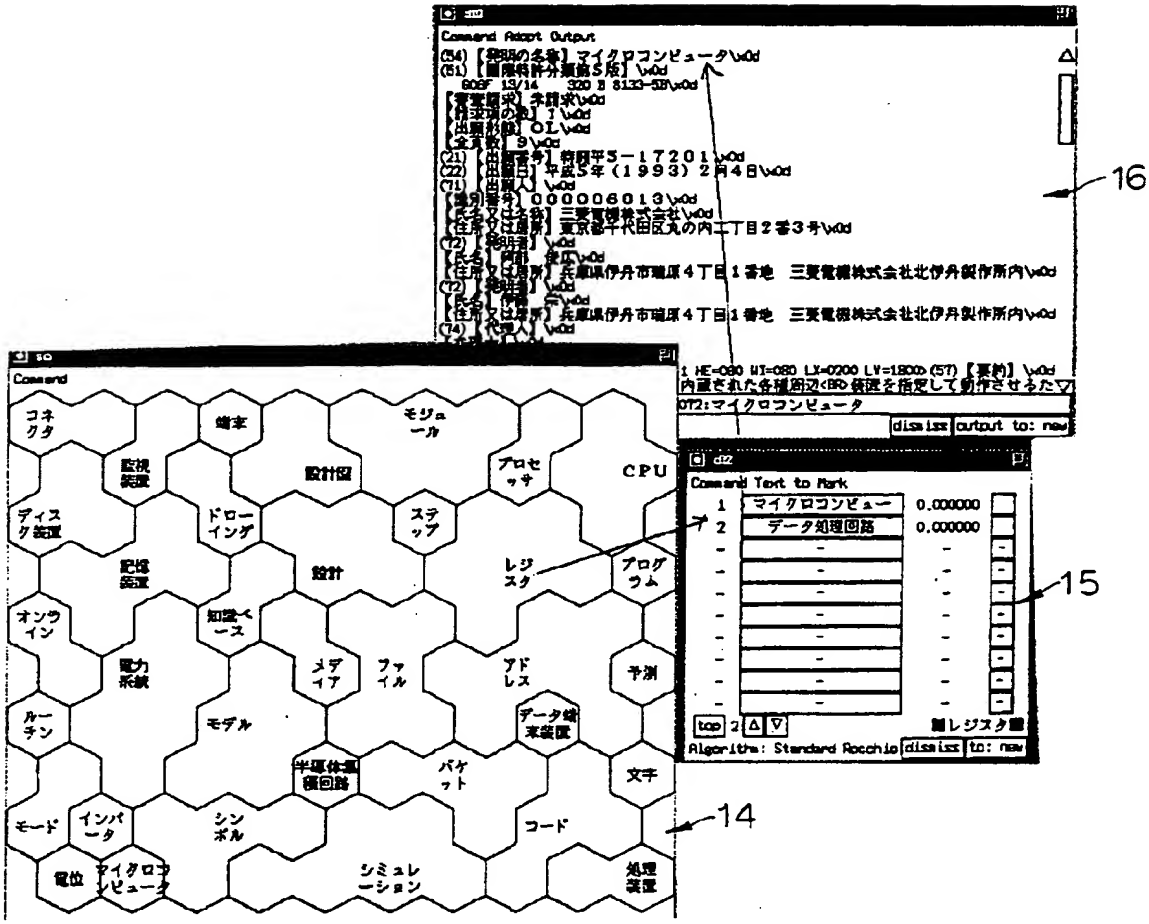
【図１１】


選択
11
12
13
情報空間可視化部
文書選択部
文書内容
表示部
10
自動分類部
17
文書分類選択格納部
9
検索結果格納部
3
検索条件
入力部
検索条件
2
情報検索部
4
検索結果
表示部
1
文書データベース

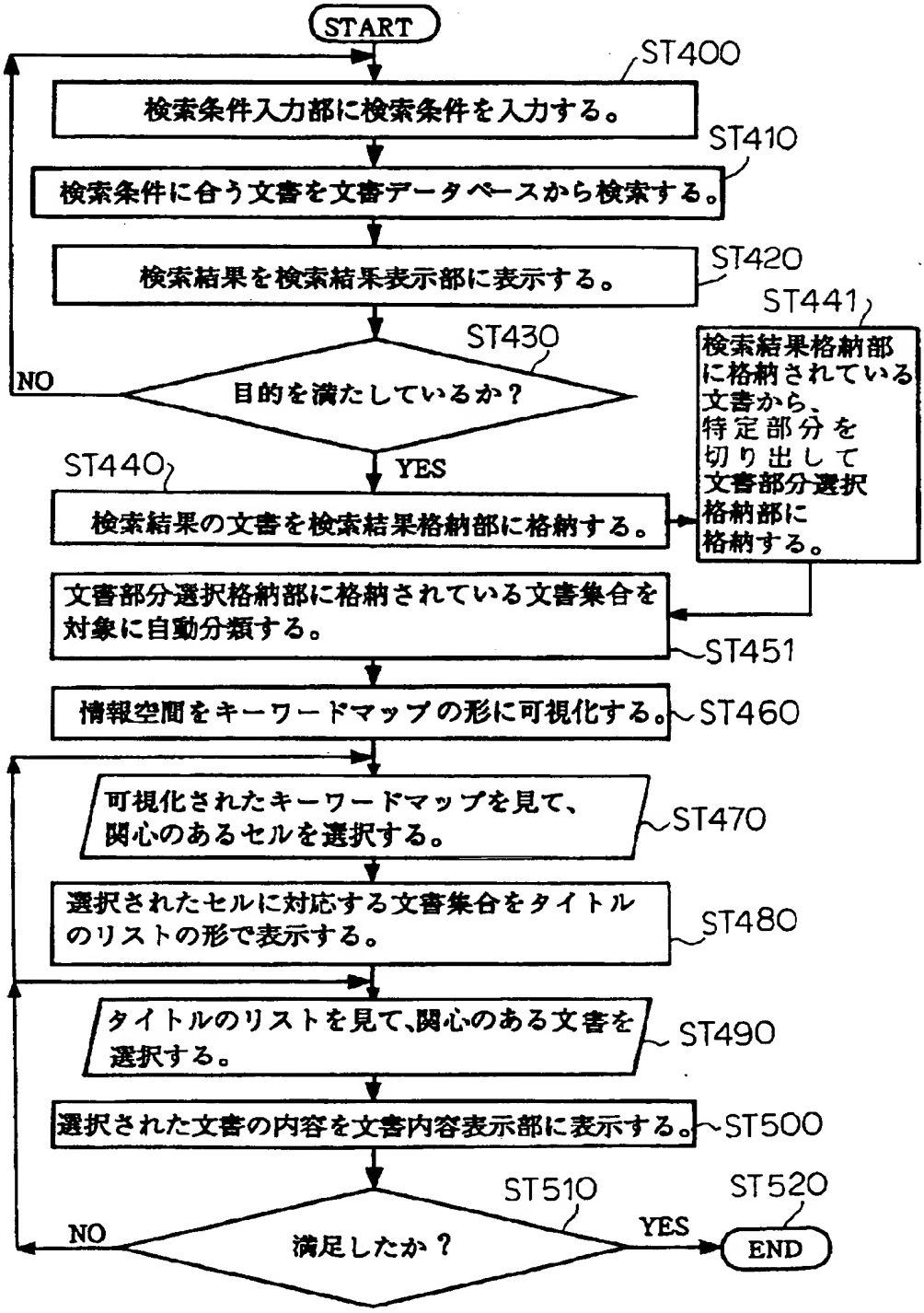
【図９】


START
ST400
検索条件入力部に検索条件を入力する。
ST410
検索条件に合う文書を文書データベースから検索する。
ST420
検索結果を検索結果表示部に表示する。
ST430
目的を満たしているか？
NO
YES
ST440
検索結果の文書を検索結果格納部に格納する。
ST450
検索結果格納部に格納されている文書集合を対象に自動分類する。
ST460
情報空間をキーワードマップの形に可視化する。
ST470
可視化されたキーワードマップを見て、関心のあるセルを選択する。
ST480
選択されたセルに対応する文書集合をタイトルのリストの形で表示する。
ST490
タイトルのリストを見て、関心のある文書を選択する。
ST500
選択された文書の内容を文書内容表示部に表示する。
ST510
満足したか？
NO
ST520
END

【図１０】


Command Accpt Output
(54)【発明の名称】マイクロコンピュータ\x0d
(51)【国際特許分類第５版】\x0d
(21)【出願番号】特願平５－１７２０１\x0d
(22)【出願日】平成５年（１９９３）２月４日\x0d
(71)【出願人】\x0d
【識別番号】０００００６０１３\x0d
【住所又は居所】東京都千代田区丸の内二丁目２番３号\x0d
【住所又は居所】兵庫県伊丹市瑞原４丁目１番地　三菱電機株式会社北伊丹製作所内\x0d
【住所又は居所】兵庫県伊丹市瑞原４丁目１番地　三菱電機株式会社北伊丹製作所内\x0d
内蔵された各種周辺<BR>装置を指定して動作させる\x0d
DT2:マイクロコンピュータ
dississ
output to: new
Command
コネクタ
端末
モジュール
監視装置
設計図
プロセッサ
CPU
ディスク装置
ドローイング
ステップ
記憶装置
設計
レジスタ
プログラム
オンライン
知識ベース
電力系統
メディア
ファイル
アドレス
予測
ルーチン
モデル
データ端末装置
半導体集積回路
バケット
文字
モード
インバータ
シンボル
コード
電位
マイクロコンピュータ
シミュレーション
処理装置
dt2
Command Text to Mark
1 マイクロコンピュー 0.000000
2 データ処理回路 0.000000
top
Algorithm: Standard Rocchio
dississ
to: new
16
15
14

【図１２】


START
ST400
検索条件入力部に検索条件を入力する。
ST410
検索条件に合う文書を文書データベースから検索する。
ST420
検索結果を検索結果表示部に表示する。
ST430
目的を満たしているか？
NO
YES
ST441
検索結果格納部に格納されている文書から、特定部分を切り出して文書部分選択格納部に格納する。
ST440
検索結果の文書を検索結果格納部に格納する。
文書部分選択格納部に格納されている文書集合を対象に自動分類する。
ST451
情報空間をキーワードマップの形に可視化する。
ST460
可視化されたキーワードマップを見て、関心のあるセルを選択する。
ST470
選択されたセルに対応する文書集合をタイトルのリストの形で表示する。
ST480
タイトルのリストを見て、関心のある文書を選択する。
ST490
選択された文書の内容を文書内容表示部に表示する。
ST500
ST510
満足したか？
NO
YES
ST520
END